

DOCUMENT RESUME

ED 229 396

TM 830 286

AUTHOR McCaig, Roger A.
TITLE How to (and How Not to) Conduct an Assessment of Student Writing.
PUB DATE [82]
NOTE 26p.
PUB TYPE Guides - Non-Classroom Use (055)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Administrators; Educational Researchers; Elementary Secondary Education; *Evaluation Methods; Holistic Evaluation; Measurement Objectives; *Models; *Student Evaluation; Test Theory; *Writing Evaluation
IDENTIFIERS Primary Trait Scoring

ABSTRACT

This paper was prepared for school officials and researchers who plan to conduct an assessment of student writing but have limited field experience with this activity. The paper identifies twelve critical questions assessors should consider, and it explores issues involved in reaching a decision about each from the perspectives of measurement theory, curriculum and instruction, language theory, and the politics of testing. Recommendations are included with supporting rationale. The key questions include choosing a mode of discourse, establishing evaluation criteria, deciding upon a rating method, selecting and training judges, planning for reliability, and using the findings. (Author)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED229396

HOW TO (AND HOW NOT TO) CONDUCT
AN ASSESSMENT OF STUDENT WRITING

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

H. Karba

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Dr. Roger A. McCaig, Director
Office of Research and Development
Grosse Pointe Public School System
389 St. Clair
Grosse Pointe, MI 48230

7M 820 286

HOW TO (AND HOW NOT TO) CONDUCT AN ASSESSMENT OF STUDENT WRITING

The inability of students to compose an intelligible, coherent passage of written English is a national disgrace and a source of outrage in communities throughout the country. Claims about the severity of the problem have been delivered with great passion by numerous personages including university officials, editorial commentators, spokespersons for self-appointed community groups, and candidates for local Boards of Education. The case for the defense, often accompanied by counterclaims about the inadequacy of funding for education, the unwillingness of parents to supervise their own children, and the deleterious effects of television, is typically delivered by the local English Department, school administrators, and officials of the teacher union. The boldness of the claim makers on both sides may be accounted for, at least in part, by the fact that none has any fear of refutation because of the absence of data to substantiate any of the claims.

This information void is paradoxical in a world virtually awash in a sea of test scores. The lay public is deluged with reports from national assessment, countless federal programs, state assessment, high school proficiency examinations, the College Entrance Examination Board, standardized testing, and other kinds of testing *ad infinitum*. Professional educators, whose business it is to use such information for the betterment of education, cannot even assimilate all the data, much less make sense of it for the improvement of teaching or the formulation of educational policy.

Yet, amid this flood of test scores, almost no useful information at all is available about student achievement in writing. No one knows the proportion of students in a given grade who can express themselves with clarity and coherence

appropriate for their age. No one can determine the relative ranking of the writing ability of an individual student or of a group of students in the same grade throughout the nation. No one knows the relationship between writing ability and any other dimension of human development or learning. The reason for this information void is plain. No one knows any of these things because there is no generally accepted system to collect data about student achievement in writing.

This article is a primer for those who have the courage to search for answers to questions about the ability of students to compose written English, whatever they turn out to be, rather than to indulge in politics, to rely on testimony, to form generalizations from a few samples of extremely poor or extremely proficient student writing. The only way to find such answers, of course, is to conduct an assessment of actual student writing. However, the process involves a number of complex issues which are not always evident to first-time assessment coordinators. An assessment plan which is incomplete or poorly conceived may produce findings which can be challenged and even dismissed as meaningless by critics who can document flaws in the process. This primer is intended as a catalog of the critical decisions assessment coordinators need to consider, along with some recommended courses of action. These recommendations involve lessons learned from the hardest of all teachers, nine years of experience in conducting an every-student assessment of developed abilities in writing for an entire school district.

Decision 1 ... Should students be permitted to write about anything they want to or should they be constrained to a specific mode of discourse? If constrained, what mode of discourse is appropriate? How does one determine appropriateness?

Option A ... Permit students to choose any topic they want to.

Option B ... Restrict the assessment to a specific mode of discourse and to a set of given topics.

Option B1 ... Choose a mode of discourse consistent with a curriculum emphasis or with communication theory.

Option B2 ... Choose a mode of discourse appropriate for the cognitive development of the students in the assessment group.

Option A, though appealing to educators who advocate theories of instruction which emphasize creativity and freedom of expression, turns out to be unintentionally unfair to students and unreliable as an assessment strategy. Research has clearly demonstrated that the quality of writing is, in part, a function of the type of writing attempted. Students who choose to write narratives will tend to receive higher ratings than they would have received if they had chosen explanatory discourse. The order of ideas in narration tends to be self-evident as in chronological narration or free as in stream of consciousness. Moreover, narrative discourse tends to provide more opportunities to student writers to display certain higher order composing abilities such as the use of non-literal language. The order of ideas in explanatory discourse, on the other hand, is not natural or self-evident. The writer must be able to plan a logical arrangement of available ideas and information about a subject as well as to compose clear, interesting sentences. Researchers and assessment coordinators, therefore, must control for this critical variable by organizing the assessment plan so that the task is the same or very similar for all students.

Choosing an appropriate mode of discourse depends upon a number of theoretical and educational positions held by the group planning the assessment. Below is an example of an assessment plan which relies both upon knowledge of child growth and development and upon curriculum emphasis. It happens to be the answer arrived at in Grosse Pointe, but groups which hold a different set of beliefs would develop a different assessment plan.

Grades 1-3

If one holds a theory of learning which uses the natural activity of children as a springboard for expressing and extending abilities, the ideal assessment mode for children in the primary grades is story telling. In Grosse Pointe the mode of discourse for assessing written expression in grades 1-3 is the literary mode. Topics are chosen which elicit children's versions of children's literature. This strategy encourages children to display forms and styles of language which they have internalized from reading and from listening to children's literature, structures which are common in written expression but occur rarely in oral language. Children in the primary grades produce their very best writing when they are guided to compose little stories in which they attribute human qualities to non-human, quasi-human, and imaginary characters.

Sample topics: The Lonely Bird, The Magic Umbrella, The Bear Who Overslept

Grades 4-5(6)

An appropriate mode of discourse for students in grades 4-6 is personal discourse. Students of this age have the best opportunity to compose language which uses the full range of their inner syntactic and semantic power when they are not constrained by the higher order cognitive problems involved in organizing a body of information which can be expressed in virtually unlimited ways.

Furthermore, since most children in these grades are still in a developmental stage in which cognitive operations are concrete rather than abstract, they find it difficult to say much at all about some subject they have not personally experienced or imagined. They may write three sentences about the Civil War or about Egypt and quit writing.

Sample topics: King/Queen for a Day, My Space Age School, The Day I Could Read Everyone's Mind

Grades 7-12

Explanatory discourse is appropriate for secondary school students, not because their best writing occurs when they compose in this mode but because it is the mode of expression necessary in our society for conducting social and business affairs and because most students have developed cognitively to a stage in which the expectation for other-centeredness in writing is neither unfair nor inappropriate. Topics selected for the middle school years should be transitional between personal and explanatory discourse. That is, although students are expected to write about a subject and to arrange the ideas in the form of a school composition, the major referents are personal experience. Topics for high school should call for personal opinion, reflection, and possibly argumentation. In all cases, choosing appropriate topics is one of the most important aspects of the assessment plan.

Sample topics for middle school: I certainly have some strange neighbors,
Parents/teachers expect too much from kids.

Sample topics for high school: Computers are changing our lives, New attitudes about the importance of equality of opportunity for both sexes have created changes in daily living.

Decision 2 ... Should students be allowed to use dictionaries and other resources in composing the writing sample?

The answer, which is no, violates preferred teaching practice and is, therefore, sometimes offensive to teachers. But testing is not teaching. Allowing or encouraging the use of dictionaries during a writing test is wrong for two reasons. The first and most obvious is that readers cannot form judgments about students' spelling ability when dictionaries are used. Do teachers allow students to use dictionaries when students are taking a spelling test? The second reason is that using the dictionary disrupts the composing process. It is more important for students to complete a flow of thought while composing than it is to spell one of the words correctly.

The solution to this apparent problem occurs in the judging, not in having students use dictionaries. Readers are instructed to form their judgments about spelling competence from a review of the spellings of basic grade-level words and from the quality of the miscues in attempting to spell difficult words. Furthermore, students who mistakenly choose "safe" words with known spellings are likely to receive middle ratings because of the failure to use effective words. The evaluation theory, therefore, turns out to support rather than to violate good instructional theory.

Decision 3 ... How much time should be planned for the assessment?

Experience has shown that thirty minutes for grades 1-6 and a full class period, about forty-five minutes, is adequate for students in secondary school. Teachers make periodic reminders of passing time as they do when administering other tests and make appropriate guiding comments like "Bring your composition to a close, now." Students who are frustrated at not being able to finish in the allotted time -- typically, superior writers in the elementary school -- are instructed to draw a line to indicate the end of the testing period and then

to finish their story. Completion of the story almost always turns out to be irrelevant to the task of evaluating the writer. The assessment task is to make a judgment about the developed abilities of the writer, not to find all the errors in the paper, and a competent judge can arrive at a rating of elementary writing from a reading of one page almost as well as from a reading of three pages.

Decision 4 ... How many writing samples should be collected from each student?

Option A ... Collect one sample per student.

Option B ... Collect more than one sample per student.

This element of the plan is crucially important, though easy to overlook. First-time assessment coordinators may not be aware that individual scores on writing tests tend to be more variable than scores on reading, math, and spelling tests. The explanation for this apparent anomaly is clear once one has become aware that the phenomenon exists. Tests in reading, math, and spelling consist of responses to very specific, carefully controlled stimuli. The responses, in fact, are often dark little marks on multiple choice answer sheets. Writing, however, is not a multiple choice test. It is an expressive ability and, as such, is somewhat more likely to vary from time to time in quality than scores on a math test, especially for the best writers. Although no one can produce an outstanding piece of writing by lucky guesswork, superior writers sometimes produce papers that fall beneath the quality of which they are capable. Professional writers, after all, often throw away their own efforts as unsatisfactory.

Collecting and scoring only one writing sample for each student, therefore, though desirable for simplicity of administration and for minimizing the disruption of instruction, is a questionable practice. The one-paper approach is

acceptable, though not preferable, when the purpose of the assessment is a comparison of group results. This procedure is based on the likelihood that an equivalent proportion of the papers in each group would turn out to be under-assessments of the true abilities of the student writers.

Collecting only one writing sample is clearly inappropriate, however, if the findings are to be used as diagnostic statements about the needs of individual students. The single-sample approach is also inadvisable if the purpose of the assessment is to determine the percent of students who have or have not developed certain abilities. A researcher, for example, might conclude from a single-sample assessment of third grade writing that only twenty-five percent of the students have developed the ability to use non-basic sentences in the process of composing written language. The scoring of a second sample, however, might produce the finding that some students who used non-basic sentences in the first sample did not do so in the second and that some who did not do so in the first sample did use non-basic sentences in the second. A researcher with access to the findings from both samples might prefer to report that forty percent of the third graders produced non-basic sentences in at least one of the two samples or that forty percent of the third graders demonstrated the ability to produce non-basic sentences.

The difference in the evaluation theory underlying the decision about how many samples to collect is crucial. The single-sample plan results in a factual report of what students did. It assumes that performance is the same thing as competence. The multiple-sample plan represents an attempt to find out what students are able to do, not what they happen to do in a single trial. The question about how many samples to collect to insure reliable results is problematical. Experience suggests that collecting two samples is a reasonable compromise between

the competing needs of reliable evaluation and limiting the disruption of teaching. A practical way to reduce the scoring load is to familiarize teachers with the criteria and to ask them to submit only the better of the two samples for formal scoring. This procedure, while not perfect, is more than adequate. Appeals can and do result from this process but the known error rate is extremely low. In the Grosse Pointe testing program less than ten scores out of about seven thousand are challenged by students, parents, or teachers, and the original score is usually upheld even after re-testing.

Decision 5 ... What other issues should be considered to choosing topics?

The topics should provide students with an equal and fair opportunity to display their writing abilities. Assigning topics which deal with special interests or which require information not readily available to all students confounds the meaning of the ratings. An interpreter of a low score cannot determine whether a student did not develop the topic adequately because of a deficiency as a writer or because of lack of knowledge about the topic. This assessment problem can be minimized by careful planning of topics and by providing students with two or three topics to select from. The practice followed in Grosse Pointe is for a curriculum committee or a representative school faculty to work cooperatively with the testing coordinator to compose the topics.¹

¹A composite list of topics which have worked best is available from the Office of Research and Development, Grosse Pointe Public School System.

Decision 6 ... What kind of prompt should be used to start students writing?

Option A ... Plan a set of topics or titles. Two or three are adequate.

Option B ... Prepare an opening passage. Break off the passage in the middle of the second or third sentence.

The difference between these options is not crucial to the quality of the writing. The choice depends mostly on the preference of the assessment coordinators. The topic approach is more open-ended and is perhaps a more complete assessment of the total composing process from start to finish than the opening-sentence approach. The opening-sentence approach gets writers started and provides more help to slow starting writers than the topic approach. It also tends to structure the mode of writing to prevent writers from composing in other forms. For instance, assessment coordinators who want to be certain that high school students write in the expository rather than in the narrative mode may wish to use an opening-sentence prompt, trading off an assessment of the process of getting started for greater structure, clarity, and uniformity in the assessment task.

Decision 7 ... Should the writing sample be prepared over a period of time or should it be impromptu in one sitting?

Although it may be argued that a prepared writing sample corresponds more closely to the kind of writing required in real life situations than an impromptu writing sample, the latter strategy is more appropriate for two compelling reasons. Developing uniform conditions for composing prepared writing involves insuperable problems in attempting to control the amount of composing time, the use of reference materials, and the authenticity of authorship. The assessment of impromptu writing is preferable to the collection of prepared writing for

theoretical as well as practical reasons. Impromptu writing characteristically contains miscues which lead to important insights about the composing processes used by the writer. Final copy does not contain the cross-outs, insertions, and other revisions characteristic of impromptu writing. The use of final copy is most appropriate when the purpose of the evaluation is to rate the quality of the papers as in a writing contest or for typical school assignments. Final copy provides information about product, not process.

Since the purpose an assessment is presumably to collect information about the composing abilities of student writers, the use of impromptu writing is appropriate as well as necessary for the integrity of the data.

Decision 8 ... What criteria should be used to judge the writing samples?

- Option A ... Review the literature and select one of the numerous evaluation scales reported therein.
- Option B ... Convene a curriculum committee and develop criteria acceptable to the local staff.
- Option C ... Select a limited set of primary traits to be evaluated.
- Option D ... Use the evaluation model developed and validated in Grosse Pointe.

The literature contains more evaluation scales than anyone can read and assimilate, much less use, dating all the way back to the first in print, Thorndike's 1911 scale. Some use evaluation topics, others use examples of student writing as models, and still others consist of guiding commentaries. Articles explain how and why teachers can, can't, should, and shouldn't evaluate writing and how and why they don't, can't, and shouldn't agree when they do evaluate writing. But there is almost no exception in the literature to the traditional practice of positing evaluation criteria by *a priori* argument or by no argument at all.

The typical evaluation scale suffers from three major flaws. One is a lack of definition of the topics. The most time-honored of all evaluation scales or topics segments the process of composing into two elements, usually referred to as ideas and mechanics. However, an attempt to determine exactly what is meant by these labels raises a number of perplexing questions. What does the topic "ideas" really mean? Does it refer to some abstract notion about the quality of the ideas independent of the language which conveys the thought, or is an idea "better," in some Pope-ian sense, if it is effectively expressed? Does the topic "ideas" refer to the logic of the argument, or does it refer also to the development of the ideas, including, for instance, the use of appropriate examples? Is the choice of words used to convey the message to be regarded as an intrinsic part of the ideas themselves, or is the choice of words to be regarded as part of the category "mechanics" depending upon the accuracy and appropriateness of the word choices? And where does sentence structure fit in? If clauses are inappropriately subordinated, is the quality of the ideas diminished, or is the problem more properly regarded as a deficiency in usage and, therefore, an error in "mechanics"?

And what of the features of language generally regarded as "mechanics"? Are they easy to score because they are quantifiable? If so, how many spelling errors are allowable for each proficiency level? Is the number of misspellings a constant or should the score be related to the number of words used? And what of the difficulty of the words? Is the misspelling of some words excusable, perhaps even inconsequential because of the difficulty of the word, or is an error an error whenever it occurs? Are all misspellings to be treated the same, or are near misspellings to be scored differently than egregious blunders? No satisfactory

answers to these and other similar questions appear in the literature. Indeed, the questions are rarely even asked. Practitioners of the traditional ideas/mechanics scale assign grades or scores to each component as if the ratings actually were to represent definable characteristics in the writing and as if the evaluation system were to follow from some theory of language.

The second major flaw in the evaluation scales which proliferate in the literature and in practice in the nation's classrooms is that no one knows the relative importance of the different components which comprise the scale, but this lack of knowledge does not deter practitioners from arriving at an overall grade. For instance, once a teacher or a judge has derived a score or a rating for spelling, how much should this score "count" in arriving at an overall rating? Is it more important than punctuation? Less important? The same? Scoring systems consisting of a number of topics typically perpetrate the ultimate evaluation fraud by averaging all the indefensible scores into a single grade or rating without any evidence at all that each of the components is of equal value.

The other major flaw is that almost no empirical evidence exists to support the inclusion of any trait in an overall evaluation of writing. Even worse, it is almost as rare to see an expression of concern about the lack of such evidence or to read about the need for obtaining such evidence. Conducting such validation research presents enormous complexities since the empirical correlates of style, organization, grammatical complexity, sentence variety, and other theoretical constructs have not been adequately defined, to say nothing at all about the components of writing that ought to be relatively easy to quantify like vocabulary, spelling, and sentence sense.

Exploratory research into these questions was conducted in the validation of the Grosse Pointe model to evaluate student writing. While the scope of this research is too extensive to report in this article, some of the pertinent findings may at least be summarized. Below is a list of the traits found to be critically related to overall ratings of writing in the validation research for the evaluation model used in Grosse Pointe.¹

Grades 1-2

Presence of a complete thought
Movement of thought
Readability
A completed "story"
Nonbasic sentences
Embedding
Creativity

Grades 3-6

Completeness
Readability
Spelling
Sentence sense
Punctuation/capitalization
Sentence structure
Connecting words
Vocabulary
Generalizations
Evidence of talent
Style

Grades 7-10

Same topics as for grades 3-6
Organization
Paragraphing
Development
Insights

The traits listed above are not scored on an elaborate multi-point scale. Instead, techniques learned from primary trait analysis are applied in the annual judging of student writing in Grosse Pointe. Each topic is scored on a simple dichotomous scale: Is there evidence in the paper that the ability has been developed or does no such evidence exist? Is the skill level OK or not OK for the grade? Each judgment applies to an appropriate competence level in a ladder-type format. A sixth grade paper judged to be not competent in spelling, for instance, would be frozen at Level 3 and could not receive a higher rating

¹McCaig, Roger A. A Model for the Evaluation of Student Writing.
Grosse Pointe, Michigan: The Grosse Pointe Public School System, 1982.

regardless of the presence of other traits. A Level 4 paper would be competent in all basic skills such as completeness, readability, and sentence sense but would not contain good sentence structure, good vocabulary, and generalizations. Competence in spelling determines whether a paper receives a rating of 3 or 5 but has nothing to do with deciding whether a paper should receive a rating of 6 or 7. The way the criteria work in the evaluation system is a direct outcome of findings in the validation research.

Given the present state of the art of evaluation, assessment coordinators would be advised to undertake primary trait analysis with confidence but to approach overall evaluation with more caution. A step in the direction of overall evaluation is to conduct a cluster analysis of traits, that is, to tally and report the percent of students who compose papers exhibiting three traits, not just one. Another alternative, short of developing one's own evaluation criteria, is to use the criteria developed in the validation research in Grosse Pointe, an approach a number of school districts and educational agencies have taken.¹

Decision 9 ... Who should judge the writing samples?

- Option A ... Local teachers, using unguided scoring.
- Option B ... Local teachers, using specific criteria.
- Option C ... Area residents who are former teachers, selected by try-out and using specific criteria.
- Option D ... A contracted service which uses acceptable criteria and which reports information about the reliability of the scoring.

Selecting the judges is perhaps the trickiest of the practical issues to be resolved in conducting an assessment of writing. The problems involved are easy to overlook because most first-time assessment coordinators approach the planning

¹
Ibid.

with two incorrect assumptions: (1) All teachers are competent judges of student writing and (2) all teachers can be trained to use specific criteria for evaluating writing.

Diederich, in his classic work on measurement in English, analyzed the ratings assigned by judges for the College Entrance Examination Board and documented what every student knows from personal experience, *viz.*, that teachers assign widely differing grades to the same set of papers.

... out of the 300 essays graded, 101 received every grade from 1 to 9; 94 percent received either seven, eight, or nine different grades; and no essay received less than five different grades from the fifty-three readers.¹

In another passage he discusses the difficulty of increasing the level of agreement even when communication within the faculty is improved, when common standards are developed, and when ratings are shared.

Even after working with an English staff for some time, I have rarely been able to boost the average correlation between pairs of readers above .50, and other examiners tell me that this is about what they get.²

This dismal picture of the state of the art of evaluating writing need not be the case, however. The evaluation of student writing collected as part of spring achievement testing in grades 1-10 in Grosse Pointe has proven to be an extremely reliable process.³ In actual practice, judges whose ratings turn out to have a reliability coefficient of less than .70 when compared to the final ratings are not invited to return as judges the next year. The result is that

¹Paul B. Diederich, Measuring Growth in English (Urbana, Illinois: National Council of Teachers of English, 1974), p. 6.

²Ibid., p. 33.

³Roger A. McCaig, "The Reliability of the Judging of the Writing System" (unpublished research report, Grosse Pointe Public School System, June, 1980).

over the years the nucleus of a highly effective judging team has been assembled, three judges for each grade. The success of the judging is due primarily to two essential components in the process.

1. Judges are selected on the basis of an actual try-out. Applicants for a vacancy are asked to attend a thirty-minute briefing session and then to score ten papers with known ratings, each paper selected to be representative of a key evaluation problem. Applicants with the best try-out ratings are selected as new judges.

The need for such a screening process became sadly evident during the early years of the current program when it was discovered that not all teachers or former teachers can become good judges, even with adequate training. Some were simply found not to be knowledgeable analysts of written language. Others could not free themselves from deeply ingrained attitudes and rules even though they often professed different beliefs. Still others appeared blind to a clever turn of phrase or to the use of irony. These shortcomings are not correctible by training but can be avoided by simply not selecting such people to serve as judges. The implications of this problem for teaching are quite another matter and are dismaying to contemplate.

2. Judges selected for their high potential are trained. The training techniques are well-known and have been in use for years in other contexts.
 - a. Judges are provided with specific criteria and specimen papers.
 - b. First, trainees rate papers independently and record their ratings.

- c. Then they compare their ratings with those given by other trainees in groups of three or four. Each trainee supports his or her rating and listens to the reasons given by others.
- d. Each group develops a consensus rating.
- e. The "right answers" are given. Trainees are helped to determine for each paper where their ratings were accurate, too high, or too low and to determine what traits they have overlooked, under-rated, or over-rated.
- f. The process is repeated as needed with different papers.

Decision 10 ... How many times should each paper be rated?

- Option A ... Have each paper evaluated by one judge.
- Option B ... Have each paper evaluated by two judges and average the ratings if they differ.
- Option C ... Plan for third readings or conferences whenever the first two readings differ.

This assessment decision is closely related to others. If, for instance, the evaluation job is relatively complex (cluster analysis of traits) and if the evaluators are not selected by screening (volunteer teachers), then a single scoring of papers is likely to produce results which are little better than random numbers. If, on the other hand, the evaluation task is relatively simple (primary trait analysis) and if the judges have been screened and trained, an assessment coordinator can take some confidence in single scorings depending upon the intended uses of the information, provided that the same judge does all the scoring. This assessment procedure is relatively safe, for instance, when the purpose of the assessment is to compare the achievement of different groups.

It is unlikely, however, that the ratings derived from a single reading will be reliable for individual students. The single-scoring procedure is also unwise when the purpose of the assessment is to determine the percent of students who have developed certain abilities. What if the single rater consistently under-scores or over-scores the papers?

Below is a summary of the options for judging appropriate for various assessment purposes:

1. Conduct third readings or conferences to resolve differences in the first two readings.

This plan may cost up to 2.5 times as much as a single reading, but, in terms of acceptability of the findings, is well worth the additional commitment. A first-time assessment of student writing is often a controversial undertaking, and when critics can cast doubt upon the meaning of the findings by locating mis-rated papers, the findings of the entire project may become suspect.

Another reason for conducting the most thorough assessment possible if conducting one at all is that the findings reported will not only be useful for research purposes but also can have enormous potential for immediate use in the classroom. In the Grosse Pointe system of evaluating writing, for instance, each writing score translates directly into a set of diagnostic statements about the developed abilities each student has acquired and about the needs for further teaching.

2. Conduct two readings for each paper and average any differences.

This procedure, which is used in assessing the CEEB writing samples, falls short of the first procedure and isn't really much better than a single reading. The possibility always exists that the first reading produced

the true score and that the average resulting from the second reading produced a mis-rating. The main purpose of collecting a second rating ought to be to confirm the first evaluation or to indicate that more analysis is necessary. Presumably, when ratings differ, one is right and one is wrong. The true score is one of the two ratings, not a mid-point between the two. In the Grosse Pointe system after all the papers have received confirming ratings and have been sorted into levels, a super judge reviews the entire batch to locate any papers which may have received the same incorrect rating two times.

3. Conduct one reading of each paper.

This procedure is not recommended except when severe constraints of time or money obtain. If the procedure must be used, it can be strengthened by sorting the papers into levels according to the ratings and by having a super judge review the entire batch. It is easier to spot a mis-rated paper when it shows up in a batch of papers all of which have the same rating than it is to determine a proper rating starting from point zero.

Decision 11 ... What method of scoring should be used?

- Option A ... Holistic
- Option B ... Primary trait
- Option C ... Combinations

A discussion of the issues involved in these options is scattered throughout this article and need not be repeated here. The decision is highlighted only

because choosing a method of scoring is an important step in the process of planning an assessment. Cooper¹ and Lloyd-Jones² have written the definitive pieces on this topic for those interested in a comprehensive discussion of the issues.

Why conduct an assessment of writing?

The possibilities for collecting information about student writing to use in improving educational services are unlimited. Below are some examples of the uses of such information in Grosse Pointe and the possibilities for any other school system which commits itself to the effort.

1. Issues about the teaching of writing are resolved or reduced.

Teachers in Grosse Pointe who used to argue that students can and should learn to write by doing workbooks and by learning grammar no longer do so. The issue is dead. Teachers who insisted on using such approaches found that at the end of the year their students could not write as well as students taught by teachers who had students do actual writing.

2. Critical tasks receive focus.

Some abilities are recognized as more critical than others and receive more emphasis than others. Different grade levels concentrate on different developmental skills. Teachers do not have to teach everything there is to learn about writing in one year.

¹Cooper, Charles R. "Holistic Evaluation of Writing." Evaluating Writing, edited by Charles R. Cooper and Lee Odell. Urbana, Illinois: National Council of Teachers of English, 1977.

²Lloyd-Jones, Richard. "Primary Trait Scoring." Evaluating Writing. Edited by Charles R. Cooper and Lee Odell. Urbana, Illinois: National Council of Teachers of English, 1977.

3. Standards and expectations are communicated to students and parents.

Every parent of an elementary school student receives an annual report of attainment in writing in a conference along with an interpretation guide. The same information is disseminated as widely as possible to the parents of secondary school students.

4. Achievement is monitored.

Attainment in writing is reported by classroom, by grade in school, and by grade in district. Grade-to-grade gain is reported for three levels of school ability to determine the growth of the writing abilities of fast, average, and slow learners.

5. Achievement is improved.

Pre- and post-test data are collected to evaluate achievement improvement projects.

6. Impact programs are conducted.

Underachievers in writing are identified and impacted in pull-out programs or by other methods. Students with disabilities in written expression can be identified for intervention by special education services.

7. Diagnostic information is provided.

Teachers and counselors receive entry scores for writing at the beginning of each year.

8. The curriculum is evaluated and modified.

The effectiveness of Writing Workshop, a required semester course in grade 10, was evaluated.

9. Research is conducted.

Examples of local research, all with clear implications for teaching, include the relationship between ability, writing, and reading; the strategies used by first graders in bringing their stories to a close; the characteristics of second graders who can compose with embedding; the relative difficulty of acquiring different writing abilities; the sequence in which writing abilities develop; and the percent of students who can be expected to attain certain writing skills in different grades.

The outcomes are clearly worth all the effort, and the entire process begins with a first brave step, conducting an assessment of student writing. Taking this first step sets in motion a chain of events which leads in directions limited only by the imagination and commitment of the faculty and administration.

HOW TO (AND HOW NOT TO) CONDUCT AN ASSESSMENT OF STUDENT WRITING

Abstract

This paper was prepared for school officials and researchers who plan to conduct an assessment of student writing but have limited field experience with this activity. The paper identifies twelve critical questions assessors should consider, and it explores issues involved in reaching a decision about each from the perspectives of measurement theory, curriculum and instruction, language theory, and the politics of testing. Recommendations are included with supporting rationale. The key questions include choosing a mode of discourse, establishing evaluation criteria, deciding upon a rating method, selecting and training judges, planning for reliability, and using the findings.